“THEME ARTICLE”, “FEATURE ARTICLE”, or “COLUMN” goes here: The theme topic or column/department name goes after the colon.

Relevancy Identification Across Languages and Crisis Types

Prashant Khare
The Open University, UK

Grégoire Burel
The Open University, UK

Harith Alani
The Open University, UK

Abstract

Social media plays a vital role in information sharing during disasters. Unfortunately, the overwhelming volume and variety of data generated on social media makes it challenging to sieve through such content manually and determine its relevancy. Most automated approaches to classify crisis data for relevancy are based on classic statistical features. However, such approaches do not adapt well to situations when applied on a new crisis event, or to a new language that the model was not trained on. In crisis situations, training a new model for particular crises or languages is not a viable approach. In this paper, we introduce a hybrid semantic-statistical approach for classifying data with regards to *relevancy* to a given crisis. We demonstrate how this approach outperforms the baselines in scenarios where the model is trained on one type of crisis and language, and tested on new crisis types and additional languages.

INTRODUCTION

Social media platforms have become a prime source for information in crises situations, enhancing situational awareness of the citizens and enabling humanitarian relief agencies to target their efforts better, [14]. Hurricane Harvey disaster saw over 7 million tweets in just over a month (digital.library.unt.edu/ark:/67531/metadc993940/). One recent study found that damage maps produced by the Federal Emergency Management Agency (FEMA) for hurricane Harvey missed around 46% of the damage reported on Twitter by impacted individuals, [15]. While this creates a source of potentially vital information, it is challenging to manually sieve through overwhelming

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

volumes of data, given that most of this information bear minimal or no relevance to particular crisis situations, even when crisis-specific hashtags are used, [6].

Previous approaches that focus on classifying crisis data as crisis related or not related, apply supervised or unsupervised approaches and rely on *statistical* features from text, such as *n-grams*, *text length*, *parts of speech* etc. and have a bias towards the nature of the data. For example, such models are likely to see a drop in accuracy when applied to a new crises type or a language that were not in training data. Retraining the models for a new crises type or language is a costly process, thus rendering it inadequate for crisis situations when immediate action is required.

We have previously shown that including semantic features enhances classifier performance when models are trained on one type of crises (example *floods*) and applied to another (example *earthquakes*), [7]. We also observed that machine translation and semantic features enhance classifiers' accuracy when training models on data in one language (example *English*) and applying it to data in different languages (example *Spanish*), [8]. However, what remains unknown is how such models perform when these two independent scenarios of cross-language and cross-crises come together, which was not explored in the previous works, [7,8].

In this paper, we tackle the problem when a model is trained on certain types of crises in a particular language (example earthquake posts in English), and tested on others (example floods in Italian). We create a multilingual dataset by translating the crises data into 6 different languages and then perform cross-crisis classification on multilingual data. We analyse the role of semantics and translation in the scenario where the model is trained on certain types of crises events (example floods) in a particular language (example Italian) and tested on a new type of crisis (not observed in training data) in a new language (example earthquakes in English). We evaluate the proposed approach in two scenarios: (1) when test data is not in the same language as of the training data; and (2) when test data is in the same language as of the training data (via translation). The main contributions are:

1. Build statistical-semantic classification model with semantics extracted from DBpedia.
2. We use 26 crises events to classify relevancy from crises events across 7 types (floods, typhoons, earthquakes, shooting, explosion, bombing, train crashes), and in 6 languages (English, Spanish, Italian, German, Portuguese, and French).
3. Evaluate 6 classifiers with multiple features, languages, and type of crises, resulting in a total of 1152 experiments.
4. Show that translating data to the same language combined with DBpedia semantics outperforms the baseline by over 16% in cross-lingual cross-crisis classification scenarios.

RELATED WORK

Social media posts during crises are not always useful or relevant. Olteanu and colleagues, [11] suggested categorising Tweets into *related and informative*, *related but not informative*, and *not related.* Much research focused on classification of such content using supervised machine learning methods that rely on statistical and linguistic features such as *POS tags*, *user mentions*, *hashtags*, *text length*, *etc*. These approaches range from traditional methods such as Support Vector Machines (SVM), [5] Naive Bayes, and Conditional Random Fields, [3] to deep learning using word embedding, [2].

In an attempt at domain adaptation, [4] authors analyse the performance of classifier when trained and tested on events from earthquakes and floods. They show that a classifier trained on Italian is likely to perform well on test events from Spanish, instead of English. However, the approach lacked a rigorous cross language scenario as it focused only on two types of events originating in Italian and Spanish languages. Few more domain adaptation approaches, [9,12] investigated learning from one type of crisis event and testing on a new type. However, these works consider only two types of the crises and also did not consider the aspect of language.

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

In previous work, [7] we used semantic information to build a relevance classifier that is adaptive to new types of crisis events. Adversarial training and graph embedding have also been used to study domain adaptability to different crises events, [1] although it was limited to only two crises types and one event per type. Previous work, [8] examined if semantic information and machine translation can render classifiers more adaptive to multilingual crisis data. Tuning classification tools to be adaptive to multilingual data is a complex task that requires enough data to build separate models for each language.

Some solutions explored in the literature are: (a) translating the data or resources (such as lexicons) from one language to target language and training the models, [10]; (b) training on weakly labelled data, [13]; (c) building multilingual distributed word-representation via Wikipedia, [16].

In contrast, we focus on the role of semantic knowledge, and automatic language translation, in classification of information relevancy in cross-lingual and cross-crisis scenarios.

RELEVANCY IDENTIFICATION ACROSS LANGUAGE AND CRISIS TYPES

We aim to (1) study the accuracy of supervised binary classification models designed for identifying crisis-related Tweets, when the type of crises and the languages used during the training phase diverge from those observed when using the models (example training on floods in English and applying to earthquakes in French), and (2) study the impact of semantic features and machine translation for alleviating crisis-type and language bias during training and usage.

For this, we train a binary classification model on certain type of crises in a particular language and evaluate it on new crises types in new languages. The proposed approach used for building and evaluating the different models used in this paper consists of 5 phases as shown on Fig. 1:

1. *Input Data and Preprocessing*: An annotated dataset containing multiple crises types is processed for alleviating training and evaluation bias.
2. *Training/Evaluation Sets Generation*: The datasets are divided into training and evaluation datasets in order to evaluate the models in a cross-crisis-type and cross-language situations.
3. *Feature Engineering*: This phase builds the statistical and semantic features that are used by the binary classifier.
4. *Model Selection and Training*: A binary classifier is trained using *crisis-related* and *not-related* annotations.
5. *Model Usage and Evaluation*: Model is evaluated on the held-out data. Depending on the approach, the language of the evaluation documents may be reconciled with the training language using machine translation.

Input Data and Pre-processing

To train cross-lingual and cross-type binary classifiers, we require multiple mono-lingual and mono-crisis-type datasets so a particular crisis-type and language can be used as training data and other crises-types and languages can be used for evaluation purposes. Twitter datasets are often composed of duplicates (example retweets) and multilingual posts that need to be sorted before using them for training a cross-type and cross-language classification model. Multilingual data in a training data may invalidate cross-lingual settings of the envisioned experiments. Finally, it is important that a binary classifier has a balanced amount of positive and negative samples during training to avoid bias towards a specific class.

We identify duplicated tweets by matching them, one by one, after removing user-handles (i.e., '@' mentions), special characters, and URLs. If the strings match, we discard the new one. With regards to language, we perform an 'identify and translate' language normalisation approach where we first use automatic methods for identifying the language of a tweet and then use machine translation for generating monolingual versions of crisis-types datasets.

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

We determine language via 3 language detection APIs: detectlanguage (detectlanguage.com), langdetect (pypi.org/project/langdetect/), and TextBlob (textblob.readthedocs.io/en/dev/) and label the language of each tweet with what is agreed by at least 2 of the APIs. More than 30 languages were found in the dataset with English (en), Italian (it), Spanish (es), and Portuguese (pt) representing nearly 93% of the data. We create a multilingual dataset for 6 languages: English (en), Italian (it), Spanish (es), French (fr), German (de), and Portuguese (pt) using Google Translation API (Neural Machine Translation System), for its superior accuracy, [17]. Each tweet is translated to the other 5 languages. If it is not in one of the 6 chosen languages, then it is translated to all 6 languages. As a result, each annotated tweet is available in 6 different languages. This results in multiple mono-lingual and mono-crisis-types datasets.

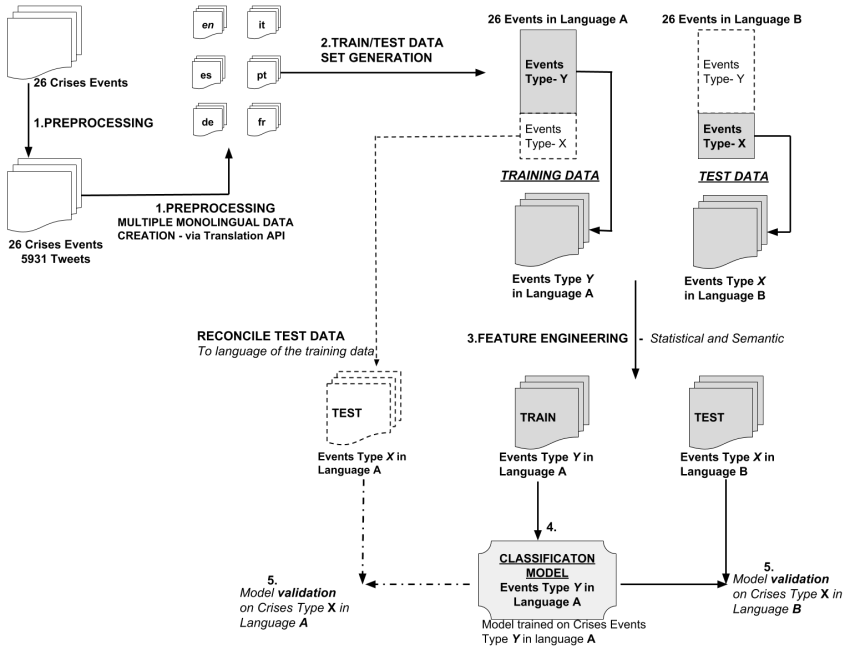

Figure 1. Pipeline for relevancy identification across language and crisis types

Training & Evaluation Sets

Since now we have monolingual datasets for multiple languages, we categorically create the train and test data. As we aim to create the training data in a certain language and for certain crises types, we choose a monolingual dataset in one of the languages and only select the crises events which are not of the *types* we aim to create the test data in. Next, to create a test data in a different language, we take another monolingual dataset in a language different than the language of the training data, and specifically select the crises events we aim to evaluate the model on. These test events do not exist in the training data.

Feature Engineering

We define two feature types: statistical and semantic features. Statistical features, the quantified statistical and linguistic properties of text, are considered as the baseline approach in our experiments. Whereas semantic features contain named entities and associated semantic information from knowledge graphs.

Statistical Features

For each post we extract the statistical features, same as in previous works, [7-8]:

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

- Number of - nouns, verbs, pronouns, words, hashtags.
- Tweet length and Unigrams

We chose spaCy (www.spacy.io) library to extract Part of Speech (POS) features. The data is tokenised to unigrams with the regexp tokenizer in NLTK (www.nltk.org). Stopwords are removed through a dedicated list of words (raw.githubusercontent.com/6/stopwords-json/master/stopwords-all.json). Furthermore, TF-IDF vector normalisation is applied over the unigrams to weigh the tokens in accordance with their relative importance within the dataset, and represent the data in the vector space.

Semantic Features

Semantic features aim to represent information in crisis data in a more generic form across languages and crisis types. These features broaden the context of documents by making them less crisis-specific, thereby reducing the problem of data scarcity.

We use Babelfy (babelfy.org/) service for the NER task, and for extracting the semantic features we use DBpedia (wiki.dbpedia.org/) as the knowledge base. Given the multilingual nature of knowledge bases such as DBpedia, some of the semantics such as labels can be extracted in English regardless of the language of the post, thus bringing multilingual data closer contextually via the added semantic vocabulary. Generalization of semantics in one language also reduces potential data sparsity resulting from varying morphological forms of entities across languages.

We extract DBpedia properties as the semantic features. The NER service Babelfy returns a DBpedia URI for each annotated entity (if available). We query the following properties associated with the DBpedia URIs via SPARQL: dct:subject, rdfs:label (only in English), rdf:type(only of the type http://schema.org and http://dbpedia.org/ontology), dbo:city, dbp:state, dbo:state, dbp:country and dbo:country (the location properties fluctuate between dbp and dbo).

Babelfy is NER service that is built on top of multilingual knowledge base BabelNet (babelnet.org). BabelNet has a common synset for entities representation in multiple languages. For instance, *police* and *policía* (police in *Spanish*) point to the same synset (unique entity node) in BabelNet- *police*. Hence the same DBpedia URI is returned for both the terms originating in different languages. We extract the additional DBpedia semantics, to associate multiple entities across the same and different languages. For instance, *guardie di sicurezza* ('security guards' in *Italian*) and *police* have the same *subject* as *Security_guard*. As a consequence, such semantics connect concepts from different types of crisis events as well as languages.

We look at two actual tweets that originate in different languages and two different types of events (Post A is from *earthquakes* in English and Post B from *floods* in Italian), gain contextual similarity as semantics are expanded; and also the similarity that translation can provide.

Post A – "*#WorldNews! 15 feared dead and 100 people could be missing in #Guatemala after quake*"

Babelfy Entities - *Feared, dead, people, missing, quake*

DBpedia Properties - *dbr:Death, dbc:Communication, dbr:News, dbc:Geological_hazards, dbr:Earthquake*

Google Translate to *it* - *#Notizie dal mondo! 15 temuti morti e 100 persone potrebbero mancare a #Guatemala dopo il*

Post B - "*Inondazioni in Sardegna, recuperato il cadavere di un poliziotto: almeno 10 tra morti e dispersi: E' morto uno d…*"

Babelfy Entities - *Inondazioni, recuperarto, cadavere, poliziotto, morti, dispersi, morto*

DBpedia Properties - *dbc:Death, dbc:Geological_hazards, dbr:Death, dbr:flood*

Google Translate to *en* - *Floods in Sardinia, recovered the corpse of a policeman: at least 10 dead and missing: He died one d .*

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

Model Selection and Training

In previous works, [7] the appropriateness of Support Vector Machine (SVM) Linear Kernel for crisis relevancy classification was validated over RBF kernel, Polynomial kernel, and Logistic Regression. Hence, we opt for SVM with a Linear Kernel as the classification algorithm.

To evaluate the use of semantic features compared to the more traditional statistical features, we design following classification models and evaluate them separately:

- SF: This model uses only the statistical features and is the baseline.
- SFSem: This model combines statistical features with semantic features from DBpedia (labels in English, type, and other DBpedia properties).

Model Usage and Evaluation

Using the held-out data previously created, we now evaluate the trained models by calculating the precision (P), recall (R) and F_1-measure (F_1). We consider two different cases for evaluating the data: (1) the model is evaluated directly by simply generating the features associated with any evaluation document (tweet) using the different features described earlier; and (2) before generating the features, we reconcile the languages used in the evaluation documents with the language used in the training data using machine translation. In any case, the training and test data still contain different crises types.

For the first use case, we use the same model notation as the training models: SF and SFSem. For the evaluation method that uses machine translation prior to the feature generation we use the following notation:

- SF^T: The model uses only statistical features, as SF, but the test data (example Portuguese) is translated to the language of the training data (example English).
- $SFSem^T$: The same model as SFSem but the test data (example German) translated to the same language as the training language (example Italian).

DATASETS

We use CrisisLexT26 dataset, [11] which is an annotated dataset consisting of 26 international crisis events of different types. Each event has 1000 labelled tweets in the following categories: *Related and Informative*, *Related and Not Informative*, *Not Related*, and *Not Applicable* and contains document in multiple languages. As shown in Table 1, the 26 events can be broadly grouped in 10 different types of crises (original language distribution shown). The categorisation is the same from previous work, [8] and is based on common understanding of the nature of the crises. For instance, *Floods* and *Typhoons* are treated as similar crisis types, since typhoons often result in floods. In this article we used all 10 event types groups for training the cross-crisis types binary classifiers.

For binary classification (i.e., *related*/*not related*) we merge the documents from all 26 crises labelled *Related and Informative* with *Related but not Informative* labels to create the *Related* class, and merged *Not Related* with *Not Applicable* to create the *Not Related* class. This allows us to train a binary relevance classifier. Then, we remove duplicates as described earlier. These processes yield 21378 unique tweets that are labelled *Related* and 2965 unique tweets labelled *Not Related*.

To avoid classification bias towards the majority class, we balance the data (whole data, and data per crisis event) by randomly under-sampling the majority class and matching the number of *Related* tweets with *Not Related* ones, across each event. This results in a final dataset with overall size of 5931 tweets (2966 Related and 2965 Not Related). And then language normalisation was performed to generate 6 versions of the monolingual datasets via Google Translation API: *en, it, es, pt, de,* and *fr*.

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

For the language reconciliation in SF^T and $SFSem^T$ models we do not need to generate new translations of the evaluation datasets since each crises-type is already translated as part of the language normalisation preprocessing task. Instead of translating the evaluation datasets we simply use the corresponding translated version of the relevant held-out data using the translated crises-types created during the preprocessing task. Overall, we had 26 crises events across 10 *types* and 6 monolingual datasets for each event. We chose to experiment on crises from *floods/typhoons*, *earthquakes*, *train crashes*, and *bombing/explosion/shooting* types.

Table 1: Event types and original language distribution (en:*English*, it:*Italian*, es:*Spanish*)

Event type	Event Instance	Event type	Event Instance
Wildfire/Bushfire	Colorado Wildfire (CWF), Australian Bush-fire (ABF) *en-99.1%, it-0%, es-0.1%, other-1.6%*	Haze	Singapore (SGR) *en-97.47%, it-0%, es-0%, other-2.53%*
Earthquake	Costa Rica (COS), Italian (ITL), Bohol (BOL), Guatemala (GAU) *en-43.6%, it-18.6%, es-30.9%, other-6.9%*	Helicopter Crash	Glasgow (GLW) *en-99.89%, it-0%, es-0.11%, other-0%*
Flood/Typhoons	Typhoon- Yolanda (TPY), Pablo (TYP) Flood- Colorado (CFL), Queensland (QFL), Al-berta (ALB), Philippines (PHF), Sardinia (SAR) *en-82%, it-12.7%, es-1.1%, other-4.2%*	Building Collapse	Savar Building (SVR) *en-86.9%, it-0.82%, es-5.19%, other-7.1%*
Terror/Shooting/Explosion	Los Angeles (LAX), Boston Bomb (BOB), West Texas (WTX) *en-95.1%, it-0.1%, es-2.1%, other-2.7%*	Location Fire	Brazil Pub (BRZ), Vene-zuela Refinery (VNZ) *en-20.3%, it-0.1%, es-45.8%, other-33.9%*
Train Crash	Spain Train (SPT), Lac Megantic (LAM) *en-47.9%, it-0.1%, es-28%, other-24%*	Meteor	Russia (RUS) *en-87.56%, it-0.64%, es-2.56%, other-9.24%*

RESULTS

We select the following crisis events and event types:

1. Train the models on rest of the crises event types except *Bombing/Shooting/Explosion* and evaluate on *LAX*, *BOB*, and *WTX*;
2. Train on rest of the crises event types except *train crash* and evaluate on *SPT* and *LAM*;

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

3. Train on rest of the crises event types except *floods* and *typhoons* and evaluate on typhoon - *TPY, TYP,* floods- *ALB, QFL, CFL, PHF,* and *SAR*;
4. Train on rest of the crises events except *earthquakes* and evaluate on earthquakes- *GAU, ITL, BOL,* and *COS*.

We already have all the events available in 6 different languages. For each pair of training data type and test data event, each time the training data is in a certain language the test data can be in the other 5 languages. This makes 30 cross-lingual cases for each event, as the training data can be in 6 different languages. Given that we have 16 test events overall, there are 480 evaluation cases in each feature model. There are 2 feature models: SF and SFSem. And the times when both training and testing data are in same language are 6 (considering translation). Hence, there are 96 such evaluation cases in each feature model. There are 2 translation models: SF^T and $SFSem^T$. Hence, effectively we performed 1152 different experiments overall. Detailed results are made available via shared repository (www.github.com/pkhare/cross_lingual_crises).

The results of performance of models and comparison over the baseline are outlined into following subsections:

Train and test in cross-lingual set ups

We measure SFSem performance over the SF baseline. Figure 2(a) shows the violin plots comparing, on average, across all 480 observations in each model. From the plots there is a clear indication that SFSem performs out rightly better over the baseline SF, with an increased overall mean and a reduced deviation in the peaks. From Table 2, we observe the mean values: SF has an avg. F_1 score of 0.556 with a standard deviation of 0.07 and SFSem has an avg. F_1 score of 0.610 with a standard deviation of 0.06. The performance of SFSem when compared to baseline *SF* is statistically significant (via 2 sample t-test) with a *p-value<0.001*.

Test data language reconciled with training data

Here we study SF^T and $SFSem^T$ performance over the baseline SF. Figure 2 (a) shows that when test data is in the same language as that of the training data, the average F_1 score increases (with or without semantics). However, adding semantics reduces the deviation. From Figure 2 (a) the highest mean is in $SFSem^T$ with a consistent distribution in comparison to rest of the models, over the baseline SF. From Table 2, we observe the mean values: SF has an average F_1 score of 0.556 with a standard deviation of 0.07, SF^T has an average F_1 score of 0.625 with 0.07 standard deviation, and $SFSem^T$ has an average F_1 score of 0.638 with a standard deviation of 0.058 (statistically significant over baseline with a *p-value<0.001*).

Overall performance across all models

Figure 2 (a) shows plots for all the models and by also taking Table 2 (showing average overall F_1 score and also across each crises type evaluated) into consideration we see that $SFSem^T$ is the best performing feature with an average F_1 score of 0.638 and an average gain of around 16.42% over the baseline SF. Without translation SFSem is the best performing feature with an average F_1 score of 0.610 with an average gain of 11.24% over the baseline SF. This shows that adding the semantics does help us alleviate the over fitting that may occur due to language and/or crisis types in training data.

We analysed the performance when the training and test data were in the same or different language (across all the tested events) as shown in Fig 2 (b) and Fig 2 (c). Note that in Fig 2 (b) and (c), when the language of training and test data is the same, it is the case of $SF^T$and $SFSem^T$ respectively across all the events in that particular language. We observe in SFT: *German* had the highest avg. F_1- 0.653 (std. dev. of 0.085) and *Italian* with lowest avg. F_1- 0.606 (std. dev. of 0.09); in $SFSem^T$ French had highest avg. F_1- 0.65 (std. dev. of 0.07) and Italian with lowest avg. F_1- 0.62 (std. dev. of 0.05).

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

The lowest standard deviation in SF^T was observed for *English*- 0.043 (average F1 score 0.63). In $SFSem^T$, *Italian* recorded an avg. of 0.62 (low standard deviation of 0.052). A much better performance was exhibited by *Spanish,* which recorded an avg. of 0.63 (low standard deviation of 0.058).

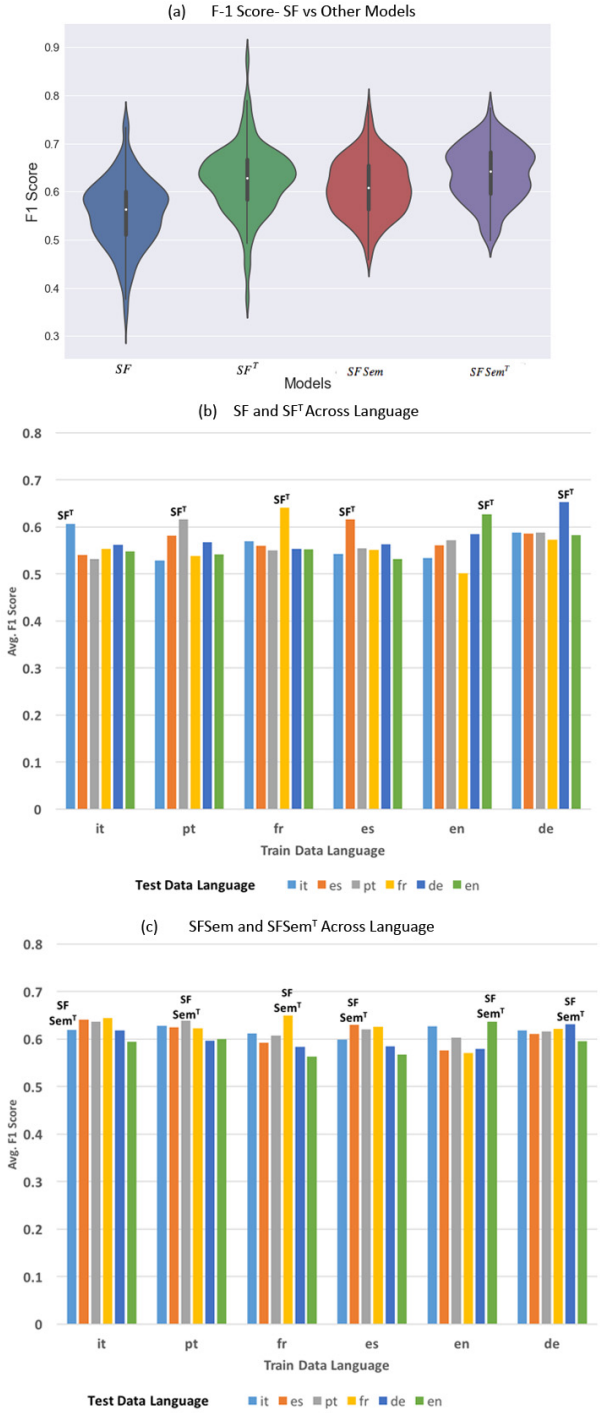

Figure 2. (a) Avg. F_1 score distribution across models (b) SF and SF^T across languages (c) SFSem and $SFSem^T$ across languages

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

Table 2: Overall performance and performance across crises types

	SF			SF^T			SFSem			$SFSem^T$		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Floods/Typhoons												
AVG.	0.618	0.583	0.551	0.698	0.66	0.643	0.684	0.648	0.628	0.711	0.683	0.672
Earthquakes												
AVG.	0.556	0.551	0.529	0.604	0.589	0.569	0.622	0.608	0.584	0.638	0.625	0.608
Bombing/Explosion/Shooting												
AVG.	0.598	0.586	0.571	0.631	0.627	0.623	0.607	0.602	0.598	0.609	0.605	0.602
Train Crash												
AVG.	0.644	0.618	0.608	0.708	0.691	0.685	0.603	0.592	0.583	0.625	0.613	0.604
Overall (across 1152 experiments)												
AVG.	0.602	0.58	0.556	0.663	0.64	0.625	0.644	0.623	0.610	0.663	0.645	0.638
STD	0.08	0.06	0.07	0.08	0.06	0.07	0.06	0.06	0.06	0.066	0.059	0.058

DISCUSSION AND FUTURE WORK

It is challenging to get a large-scale annotated data across several languages and event types; hence we simulated the scenarios by translating the original data (for each crisis event) to 6 different languages. Google Cloud allows a maximum of 10M-character translation per 100 seconds per projects (https://cloud.google.com/translate/quotas). Machine translation, much like NLP tools and semantic expansion via knowledge bases, might be less accurate or unavailable for non-European or low resourced languages. Some of the statistical features are language independent. However, our aim was to explore the feasibility of these methods in such problems. And we observed that both methods enhance the classification accuracy. The translation brings vocabulary in the same language, whereas the semantics align the context. Also, the 16 crises events were majorly across four types. Next, we aim to obtain a wider representation of crises types.

Previously, [8] we performed cross-lingual experiments with *balanced* and *unbalanced* datasets. *Unbalanced* datasets, while often representing real world data distributions, create bias towards the dominant class. While we have managed to categorically break down the events based on their types, we cannot ignore the possibility of some overlap residue. In future work, we could take similarity measurements to quantify and reduce any such remaining overlap.

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

CONCLUSIONS

For better use of social media platforms during crises events, it is crucial to efficiently determine the relevant content. In this paper, we took a wider and more realistic scope of the problem, where the type of the crisis, and the language of incoming content, vary a lot, which could have serious consequences on the validity of any given content-relevancy classification model. We were able to test various models, built on different approaches. Primarily, we explored the impact of adding semantics and automatic language translation, independently and in combinations. We showed that if translation is not feasible then the combination of statistical and DBpedia features is the best performing approach; and if translation is viable then a combination of translation, statistical and DBpedia features is the best performing approach.

REFERENCES

1. Alam, F., Joty, S. and Imran, M., 2018. Domain Adaptation with Adversarial Training and Graph Embeddings. arXiv preprint arXiv:1805.05151.
2. Burel, G., Saif, H. and Alani, H., 2017. Semantic wide and deep learning for detecting crisis-information categories on social media. ISWC, Vienna.
3. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. and Meier, P., 2013. Practical extraction of disaster-relevant information from social media. 22nd ACM WWW (World Wide Web), Rio de Janeiro
4. Imran, M., Mitra, P. and Srivastava, J., 2016. Cross-language domain adaptation for classifying crisis-related short messages. arXiv preprint arXiv:1602.05388.
5. Karimi, S., Yin, J. and Paris, C., 2013, December. Classifying microblogs for disasters. In Proceedings of the 18th Australasian Document Computing Symposium. ACM.
6. Khare, P., Fernandez, M., Alani, H., 2017. Statistical semantic classification of crisis information. In: Workshop on HSSUES at ISWC, Vienna
7. Khare, P., Burel, G., Alani, H., 2018. Classifying crises-information relevancy with semantics. In: ESWC, Crete
8. Khare, P., Burel, G., Maynard, D. and Alani, H., 2018. Cross-Lingual Classification of Crisis Data. ISWC, Monterey, US.
9. Li, H., Caragea, D., Caragea, C. and Herndon, N., 2018. Disaster response aided by tweet classification with a domain adaptation approach. Journal of Contingencies and Crisis Management
10. Mihalcea, R., Banea, C., Wiebe, J., 2007. Learning multilingual subjective language via cross-lingual projections. 45th ACL, Prague
11. Olteanu, A., Vieweg, S., Castillo, C., 2015. What to expect when the unexpected happens: Social media communications across crises. CSCW, Vancouver
12. Pedrood, B. and Purohit, H., 2018. Mining help intent on twitter during disasters via transfer learning with sparse coding. In SBP-BRiMS
13. Severyn, A., Moschitti, A., 2015. Unitn: Training deep convolutional neural network for twitter sentiment classification. International workshop on semantic evaluation. Colorado
14. Vieweg, S., Hughes, A.L., Starbird, K. and Palen, L., 2010, April. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. SIGCHI-CHI. Atlanta.
15. Villegas, C., Martinez, M. and Krause, M., 2018. Lessons from Harvey: Crisis Informatics for Urban Resilience. Kinder Institute for Urban Research.
16. Wick, M., Kanani, P., Pocock, A.C., 2016. Minimally-constrained multilingual embeddings via artificial code-switching. AAAI, Arizona.
17. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. and Klingner, J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

This is an accepted version of the article for publication in IEEE Intelligent Systems, and is not final edited version. Content may change in edited version. Citation information: DOI 10.1109/MIS.2019.2917443, IEEE Intelligent System 1541-1672 (c) 2019 IEEE.

ABOUT THE AUTHORS

Prashant Khare is a PhD student at the Knowledge Media Institute, The Open University, UK. He has previously received Master of Science in Web Technology from University of Southampton, UK. Contact him at prashant.khare@open.ac.uk

Dr. Grégoire Burel is a research associate and data scientist at the Knowledge Media Institute (Open University) involved in the COMRADES European project and the lead developer of the Crisis Event Extraction Service (CREES). Contact him at g.burel@open.ac.uk

Prof. Harith Alani is a Professor of Web Science at the Knowledge Media Institute, The Open University. He is also a member of Elsevier Advisory Panel. Contact him at h.alani@open.ac.uk